

SCALABLE DERIVATIVE SERVICES

Inventors

Sheng Liang
Oliver Chang
Hong Zhang
Abhishek Chauhan
Rajiv Mirani

SCALABLE DERIVATIVE SERVICES

Inventors

5

Sheng Liang
Oliver Chang
Hong Zhang
Abhishek Chauhan
Rajiv Mirani

10

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims priority under 35 USC §119(e) from United States Provisional Application Serial Number 60/228,904, filed on August 29, 2000 and incorporated herein by reference in its entirety. In addition, this application is related to Application Serial Number 09/687,997, filed on October 13, 2000, and which is also incorporated herein by reference in its entirety.

15

BACKGROUND OF THE INVENTION

Field of Invention

The invention relates to improved parsing and manipulation of data such as a markup language.

20

Description of the Related Art

With the increased popularity of web-hosted services and applications, portals and other Service Providers (SPs) have begun to offer hosted applications that enhance, or even replace, traditional desktop applications. For example, it is increasingly more common to compute one's taxes, pay bills, and manage daily appointments and e-mails

via the World Wide Web, instead of through traditional locally-installed software applications.

Along with this growth, there is an accompanying need for derivative services that leverage and integrate existing web-hosted services, allowing end users to access
5 data obtained from multiple services and presented in a single page or set of pages.

Building derivative services for applications hosted across various domains poses new challenges compared with building services such as search engines and shopping comparison sites. Search engines and shopping comparison sites handle web
10 pages that need not be specific to individual users and can remain static for a sizeable period of time, since content may often remain unchanged for hours, days or even weeks at a time. In contrast, derivative services for hosted applications must deal with highly dynamic and personalized web pages. For example, it would not be at all
15 desirable for old e-mail to be redelivered, or for new e-mail to be substantially delayed because of a service that was slow to update. These derivative services should therefore interact with hosted applications on behalf of the end user in real time.

To provide derivative services, there is a difficulty of having to scale the number of virtual browsers, which establish connections from a derivative services provider (DSP) to primary web servers to retrieve content on a user's behalf, to the number of concurrent users of the DSP. Traditional browsers are designed for desktop use, and
20 typically require several megabytes of memory to run. Virtual browsers are not being run on the users' machines, but instead are being executed by the DSP. As a result, content retrieved by a virtual browser may not even be displayed on any screen, but instead just passed to another system for further processing. If a separate virtual browser were instantiated on a DSP server for each end user, the resulting tax on the
25 server's processor and memory could quickly become overwhelming. However, in order to provide the kind of derivative services discussed above in which existing web-hosted services from primary servers are integrated and delivered simultaneously, there

5

is a need for multiple virtual browsers to operate simultaneously on the DSP side. This problem, i.e. the need for multiple virtual browsers and the difficulty in providing them, has been an obstacle to the growth of the derivative services area. Accordingly, what is needed is a system and method for providing scalable derivative services that avoids the processing and memory drain present in current implementations.

BRIEF SUMMARY OF THE INVENTION

The present invention efficiently parses HTML content (or other markup or content language such as DHTML, Java, etc.) by identifying data files (e.g. pages) that contain some unchanging (static) content, but still vary to some degree with each new version. Once these pages are retrieved and identified, they are parsed to form an abstract syntax tree (AST), and the pages and their associated ASTs are then cached. When a new version of a page that already is stored in the cache is retrieved, the new version of the page is compared to the stored version to determine which portions of the new version contain new content. Nodes of the AST corresponding to content that does not vary between the versions of the page are identified as static nodes; the remaining content is deemed to be dynamic content. Once the dynamic content of the page has been identified, it is parsed to form dynamic AST nodes, which may be combined with the cached static nodes to form a complete AST. Since only the portions of the new version of the page corresponding to dynamic content are reparsed, computation time and memory are thereby saved. This enables a larger number of virtual browsers, which jointly access the ASTs, to be used in a derivative server, which can use the ASTs and associated pages to deliver derivative services to many remote clients.

25

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is an illustration of one way in which derivative services are provided by a service provider to a client.

Fig. 2 is a screen shot of an HTML page displaying a real time stock quote.

Fig. 3 is an illustration of an abstract syntax tree built from the HTML page of

5 Fig. 2.

Fig. 4 is an illustration of the dynamic content present in the abstract syntax tree of Fig. 3.

Fig. 5 is a block diagram of a preferred embodiment of a system in accordance with the present invention.

10 Fig. 6 is an illustration of static nodes of an abstract syntax tree in accordance with the present invention.

Fig. 7 is an illustration of an abstract syntax tree containing static and dynamic nodes in accordance with the present invention.

15 Fig. 8 is a flowchart of the operation of a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The present invention includes a system and method for providing derivative services on a scalable basis to a large number of users simultaneously by taking advantage of the fact that often, only a small portion of a page changes each time the page is retrieved for display to a user. By only parsing those portions of a web page that are dynamic, i.e., that change from instance to instance, significant memory and CPU processing time can be saved. This allows for increased scalability of the virtual browsers and reduced memory requirements compared to present day systems, consequently enabling service to a greater number of simultaneous users.

Referring now to Fig. 1, there is shown an illustration of how derivative services are provided by a derivative service provider (DSP) 100 to a client 102.

The DSP 100 gathers content from a plurality of web servers 104 and uses a derivative server 110 to combine the content to produce a page 108 (or set of pages) for the client 102. (One client is shown, but in practice, the DSP 100 will maintain concurrent connections with thousands of clients 102, providing the described functionality for each.) The client 102 does not access each web server 104 directly, but rather obtains an assembled page 108 from the DSP 100 after sending it a request for a page. In many cases, the requested page does not exist on any one server 104, but rather is constructed by the derivative server 110 from multiple pages from different web servers 104. Each web server 104 is accessed by a virtual browser 106 belonging to the DSP 100. Since a typical browser such as Microsoft's Internet Explorer or Netscape's Navigator can easily require several megabytes of memory to execute, there is a high cost to simultaneously executing a large number of virtual browsers. For a DSP 100 to serve hundreds or thousands of clients 102 simultaneously, with each client 102 requiring a different dedicated virtual browser 106 for each site 104 forming part of that client's assembled page 108, the DSP 100 would have to have hundreds or thousands of

virtual browsers simultaneously available, with the memory and processing power available to support them.

One of the reasons the virtual browsers 106 require so much available memory is that they typically parse every page they retrieve. Parsing is a memory intensive activity because it requires the use of data structures such as tables or trees. One method of parsing involves building an abstract syntax tree (AST). To illustrate the formation of an AST, consider Fig. 2, which depicts a web page 200 that provides a real time stock quote. The page comprises a title portion 202, a time stamp 204, a ticker symbol 206, a price 208, a text entry box 210, and a "Get Quote" button 212. The HTML code for such

a page might resemble the following:

```
10 <html>
11   <head>
12     <title>Real time stock quote</title>
13   </head>
14   <body>
15     Real time stock quote <br>
16     (2:31:45.25 PM 6/28/2000)
17     <p>
18     <table border=0>
19       <tr bgcolor="#dddddd">
20         <td width=100>MSFT</td>
21         <td width=57>89.125</td>
22       </tr>
23     </table>
24     <p>
25     <form action="quote.html">
26       <input type="text" name="symbol" size="8">
27       <input type="submit" name="submit" value="Get Quote">
28     </form>
29   </body>
30 </html>
```

An AST built from this page is illustrated in Fig. 3. The AST 300 comprises various nodes, including nodes corresponding to the title banner 302, the time stamp 304, the stock ticker symbol 306, the stock price 308, and the "Get Quote" button 312. As Fig. 3 illustrates, the resulting AST 300 that is produced is large, compared to the small amount of HTML code used to build the page.

The present invention enables derivative services that overcome the excessive memory and processor requirements of existing systems by taking advantage of the fact that much of the content in a given content page remains static over time. By determining which portions of a page are static, and which are dynamic, the present invention provides the ability to parse only the dynamic content, therefore reducing the memory and processing time required by the virtual browsers, and thus increasing the number of virtual browsers that can be run concurrently by a DSP.

Returning to the stock quote page described above with respect to Fig. 2, most of the page is static, and only a small portion is dynamic. The title 202, text entry box 210, and "Get Quote" button 212 are static elements of the page 200, in that they do not typically vary each time the page 200 is retrieved. Conversely, the time stamp 204, stock ticker symbol 206, and stock price 212 are dynamic, and would be expected to vary.

Referring now to Fig. 4. there is shown the AST 300 described earlier with respect to Fig. 3. Fig. 4 illustrates that nodes 304, 306, and 308 are dynamic, while the remaining nodes, including nodes 302 and 312 are static. As can be seen from Fig. 4, the dynamic portion of the AST 300 is small compared to the overall size of the AST. Thus, it would be desirable if the entire page, containing mostly static content, did not have to be reparsed every time the page is fetched.

The present invention provides just such a solution. Referring now to Fig. 5, there is shown a block diagram of a preferred embodiment of a system in accordance with the present invention. System 500 comprises one or more virtual browsers 502, a cache 504, a comparison engine 506, a content server 508, and a token master 516. Virtual browser 502 additionally comprises an identification engine 510 and a parsing engine 512. Also shown in Fig. 5 is a client computer 514. A derivative service provider 100 may in practice have more than one system 500 in place depending on the capabilities of the hardware used by the provider 100 and the number of client computers 514 and web servers 104 that must be accessed simultaneously. The

connection between client computer 514 and system 500 is via the Internet in a preferred embodiment, but may be by direct dial-up, LAN or other network.

When the client 514 attempts to access the DSP 100 service provided by system 500, content server 508 determines which web pages must be accessed on corresponding primary web servers 104 in order to build a content page (or pages) 108 to give to the user. One or more virtual browsers 502 are then assigned to retrieve the content (HTML or other content formats) from the primary web servers 104.

When a page is retrieved by the virtual browser 502, it is identified by the identification engine 510. The identification engine 510 determines whether the identified content object is one that is currently being tracked. If the identification engine 510 determines that the content object is not one that is already being tracked, then the identification engine determines whether the object is one that should be tracked. If the object is not one that should be tracked, it is simply parsed by the parsing engine 512 and sent to the content server 508. If it is to be tracked, the object is stored in the cache 504 along with its associated AST formed by the parsing engine 512 in addition to being sent to the content server 508.

If the identification engine determines that the object is being tracked, then the currently retrieved object is a new version of the object. Accordingly, the comparison engine 506 identifies the differing content in the new version of the page, by comparing the new version with the original (stored) version. The parsing engine 512 then parses the differing content, which is then associated with the original static content and sent to the content server 508. When content server 508 receives content from the parsing engine 512, it uses the received content to form additional derivative services content for transmission to client computer 514.

The operation of system 500 is now considered in greater detail. A user of client computer 514 accesses system 500 in order to retrieve derivative services content. For example, the user may have set up an account with the DSP 100 that provides him with

e-mail, stock quotes, weather and news, each from a different original service provider (e.g. primary web server 104), but presented on one page 108 by the DSP 100. To access the DSP 100, the user initiates a connection from the client computer 514 to the content server 508 of system 500 via the Internet. Once connected, the client computer 514 sends 5 a request to system 500 to provide derivative services.

The content server 508 determines which web servers 104 contain content that must be assembled and supplied to the client computer 514. The content server 508 then sends one or more requests to the virtual browser 502 for a content object (a page). Note that while in a preferred embodiment, the requested object is an HTML object, in 10 alternative embodiments the item to be retrieved could be MIME-encoded email messages, XML pages, or other structured content.

The virtual browser 502 retrieves the requested page, and asks the identification engine 510 to identify it. Identification engine 510 attempts to identify the page according to rules stored in the identification engine 510. For example, suppose that the 15 page is CNNfn's stock quote page. A URL on that page is of the form "http://qs.cnnfn.cnn.com/tq/stockquote?symbols=MSFT", where "MSFT" is the symbol for the stock price being looked up. If a different quote, e.g. "AAPL" were retrieved, the URL might be "http://qs.cnnfn.cnn.com/tq/stockquote?symbols=AAPL".

Thus, the identification engine might store a rule that identifies URLs containing the 20 string "http://qs.cnnfn.cnn.com/tq/stockquote?symbols=" or perhaps even "http://qs.cnnfn.cnn.com/tq/stockquote?" as an appropriate key for identifying the page.

While in a preferred embodiment, the rules stored in the identification engine 25 510 for identifying keys are entered manually, alternative embodiments allow the identification engine 510 to automatically select keys by analyzing similar patterns of URLs seen over periods of time. For example, in the example above, if a URL is seen

repeatedly by the identification engine with only small changes each time, e.g. "MSFT" replaced by "AAPL" or "IBM", the identification engine 510 extracts the non-changing portion of the URL and forms a key from it. In addition to URLs, keys may include identification codes embedded in each page, or other indicia.

5 After a page has been identified, the identification engine 510 determines whether the page is already being tracked by system 500. This is done by maintaining a list of pages (indexed by key) in the cache 504. In other embodiments, the token master 516 maintains a table of keys that are being stored.

10 If the page is not already being tracked, then the identification engine 510 determines whether tracking should be initiated. Ideally, the best page candidates for tracking by system 500 are those that contain a mix of static and dynamic content. These pages should not be cached and reused in their entirety, because some data changes each time the page is retrieved. However, as in the stock quote example above, since much of the page remains constant each time, it would be wasteful to reparse the entire page 200. These characteristics make mixed pages (i.e. pages containing both dynamic and static content) well suited for tracking by system 500.

15 In a preferred embodiment, the virtual browser 502 accesses a rule database, which comprises a list of pages or keys that should be tracked. If the key for the page currently being retrieved matches a key listed in the rule database, then the page is tracked. 20 In other embodiments, the identification engine 510 determines over time whether the page is suitable for tracking. Each new page (i.e. one not recognized by the identification engine 510) is initially tracked by default. If after the page has been seen a certain number of times, e.g. five, no static content can be identified, the page is no longer tracked. In addition, if the page is not seen again within a specified period, 25 measured in either time or volume of pages retrieved, the page is expunged from the cache 504 and no longer tracked.

Referring now to Fig. 6, once it is determined that the page should be newly tracked, it is parsed by the parsing engine 512. As the page is parsed, a template/token tree 600 is built. A template/token tree is an AST modified to contain nodes and tokens. Each node of the tree 600 is initially labeled as a static node, indicating that the content stored in that node is to be treated by system 500 as static content. Node 602 is a typical static node of template/token tree 600. Once the HTML has been parsed and the template/token tree 600 built, the page and tree 600 are stored in the cache 504 until a version of the page is again retrieved by the virtual browser 502.

If the virtual browser 502 determines, on the other hand, that the page is already stored in the cache 504, the page and its associated template/token tree are loaded from the cache. System 500 does not need to parse the entire new page version, so long as it contains at least some text that is identical to the cached copy. (Note that in this description, the "version" of a page refers to the specific page as it appeared on the particular occasion on which it was retrieved. That is, if a page is retrieved at time t_0 and then retrieved again at some new time t_1 , two versions of the page have been retrieved.) In order to determine how much of the new page version to parse, comparison engine 506 examines the different versions of the page. In a preferred embodiment, the newly retrieved version of the page and the cached copy of the page are compared using a binary "diff" algorithm, which identifies the differences between the binary representation of two documents. Binary diff is used in a preferred embodiment in part because its execution time is fast compared to the speed of having to parse HTML. In other embodiments, other comparison techniques may be used, with the caveat that comparisons that take increasingly longer to make will result in decreasing performance advantages over total parsing.

Once the comparison engine 506 determines which content varies between versions of the page, i.e. between the cached version and the newly retrieved version, the template/token tree associated with the page is updated so that nodes containing

the different text are replaced in the tree by tokens 704. Tokens are requested from and assigned by the token master 516, and contain a unique ID. The virtual browser 502 forms subtrees from the nodes containing dynamic text, and additionally maintains a mapping from each token to its associated subtree. Referring now to Fig. 7, there is
5 shown an illustration of a template/token tree having nodes that contain static content (static nodes) 602 and nodes that contain dynamic content (dynamic nodes) 702. The dynamic nodes 702 of the template/token tree are replaced by tokens 704, which are mapped by the virtual browser 502 to one or more subtrees 706 containing the dynamic nodes.

10 Once system 500 replaces the dynamic content with tokens 704, content corresponding to the static nodes 602 of the template/token tree need not be reparsed each time a new version of the page is retrieved. Instead, only dynamic content is parsed. The tokens of the template/token tree are mapped by the virtual browser 502 to the subtrees containing the dynamic content associated with the newly retrieved version
15 of the page. Since system 500 in a preferred embodiment contains a large number of virtual browsers 502 operating currently to serve multiple users, each virtual browser 502 is responsible for maintaining its own mapping of tokens to dynamic subtrees.
Thus, when a token/template tree 600 is retrieved from the cache by a virtual browser, it
comprises static nodes and tokens with unique identifiers. As the virtual browser
20 retrieves and parses dynamic content, it builds dynamic subtrees and associates each unique token identifier with a mapping to a specific subtree. This mapping remains, in a preferred embodiment, for as long as the virtual browser 502 has a connection open to the primary web server 104 associated with the dynamic subtree. If, in the course of identifying non-matching content, the comparison engine 506 determines that one of the
25 template/token tree's static nodes actually contains dynamic content, the comparison engine requests a new token from the token master 516. This token replaces the dynamic node in the template/token tree, and the virtual browser creates a mapping

from the new token to a subtree containing the associated dynamic content. When the virtual browser returns the template/token tree to the cache 504, it will contain the new token assigned by the token master 516. In this way, dynamic content is not cached once it is identified as dynamic, which results in lower storage space requirements, and
5 additionally avoids data that may be sensitive, such as a user's password or financial information.

Note that if any portion of a node contains dynamic information then the entire text corresponding to that node is reparsed. For example, using the stock lookup page 200 of Fig. 2, suppose the ticker symbol being looked up in the first version is "MSFT." Then suppose that in the second version, a user has looked up the stock with ticker symbol "MOT." Since the binary diff algorithm only reports text that is different between versions of the page, the "different" text will be "OT" in the new version of the page, instead of "MOT," since both "MSFT" and "MOT" begin with the letter "M." However, since the stock ticker symbol is part of node 306, the entire node will be reparsed.
10
15

Referring now to Fig. 8, there is shown a flow chart of the operation of a preferred embodiment of the system. To begin, system 500 receives 901 a request for an HTML object. System 500 then retrieves 802 the requested HTML object, e.g. via HTTP. System 500 identifies 803 the object and determines 804 whether the object has been previously stored in the cache. If the object is new, system 500 determines 806 whether the object is one that should be tracked by the system 500. If the object is not one that should be tracked, it is sent 822 to the content server 508. If the object is to be tracked, then it is sent to the parsing engine to be parsed 808. As the object is parsed, a corresponding template/token tree is built 810, and each node in the tree is initially designated as containing static content. The tree and the associated HTML object are then stored 812 in the cache. The parsed object is then sent 822 to the content server 508 where it is assembled for presentation to an end user as derivative content.
20
25

When system 500 retrieves 802 an HTML object that is determined 804 to be one that has been previously encountered and stored, a previous edition of the text and template/token tree are retrieved 816 from the cache 504. The text of the new HTML object is then compared 818 to the cached version. Content of the new object which differs from the cached version is then parsed 819. The template/token tree 600 is then updated 820 to comprise static nodes 602, which correspond to text not found to differ between objects, and unique tokens assigned by the token master 516 replacing content that does differ between objects. The updated template/token tree and most recent version of the HTML object are then returned 820 to the cache 504.

Consideration should also be given to the realization that web sites occasionally change the underlying templates. System 500 therefore monitors the ratio of static to dynamic content in the each page document being tracked. If the ratio of static to dynamic content in a page changes drastically, e.g. if the amount of dynamic content is suddenly found to have doubled, the underlying template may have changed and system 500 simply flushes the HTML document from the cache and restarts the iterative template-building process.

As will be understood by those familiar with the art, the invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. For example, although the focus here is on generating template/token trees from HTML pages, those of skill in the art will recognize that system 500 can be extrapolated to effectively parsing many kinds of serialized structured content. Likewise, the particular modules, engines, protocols, features, attributes, data structures, or any other aspect is not mandatory or significant, and the mechanisms that implement the invention or its features may have different names or formats.

Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.